(N)

PAT-NO: JP406215196A

DOCUMENT-IDENTIFIER: JP 06215196 A

TITLE: METHOD FOR ESTABLISHING GROUP DATA BASE IN CHARACTER RECOGNITION

PUBN-DATE: August 5, 1994

INVENTOR-INFORMATION:
NAME
TO, RAKUTEI
SEN, EIKAN
JO, EIJI
RIN, BUNBUN

ASSIGNEE-INFORMATION:

NAME	COUNTRY
IND TECHNOL RES INST	N/A

APPL-NO: JP04318780

APPL-DATE: November 27, 1992

INT-CL (IPC): G06K009/68

ABSTRACT:

PURPOSE: To reduce overlapped amounts, and to improve group applying precision by applying a pattern clustering group center to a group.

CONSTITUTION: The plural leaning patterns of one character are ununiformly divided (ST1), and the characteristics of the learning patterns are extracted (ST2). Then, the constant number of features are selected from each feature of the learning patterns, and the capability of a group is obtained (ST3). The learning patterns are divided into ω groups by a clustering method based on the selected features (ST4). This ω is an integer larger than 1. Then, the center value of the ω groups obtained by clustering the learning patterns is applied to the group as a central pattern, overlapping is obtained, and the best grouping is obtained (ST5). The clustering method includes a random clustering, ISODATA clustering, and K-mean clustering, and they are sequentially executed.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－215196

(43)公開日　平成６年(1994)８月５日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｋ　9/68		9289－5L		

審査請求　未請求　請求項の数３　ＯＬ　（全　６　頁）

(21)出願番号　特願平4－318780

(22)出願日　平成４年(1992)11月27日

(71)出願人　390023582
財団法人工業技術研究院
台湾新竹縣竹東鎮中興路四段195號

(72)発明者　屠　樂　梃
台湾台北市士林區名山里５鄰▲雨▼聲街53巷２號４樓

(72)発明者　▲せん▼　永　寛
台湾彰化縣永靖郷▲なん▼港村永社路260巷２鄰18號

(72)発明者　徐　英　士
台湾台北市北投區文林里16鄰致遠一路一段47巷１號２樓

(74)代理人　弁理士　伊東　忠彦　（外１名）

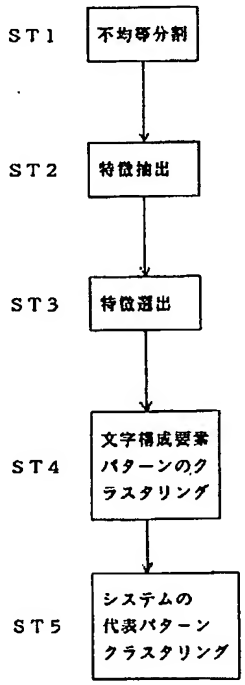
最終頁に続く

(54)【発明の名称】　文字認識におけるグループデータベース確立の方法

(57)【要約】

【目的】　文字認識において文字の代表パターンを利用してシステムとしてグループデータベースを確立する方法であって、パターンが多すぎること及び変異が大きすぎることに起因するシステムのグループデータベースの拡張（重複量の増大）を回避すると同時に、システムのデータベースの保存スペースを削減する文字認識におけるグループデータベース確立方法を提供することを目的とする。

【構成】　１つの文字について多数個の学習パターンをそれぞれ不均等分割すること、学習パターンの特徴を抽出すること、学習パターンの各特徴の中から一定数の特徴を選出して、グループを確立すること、選出された特徴に基づいて、学習パターンをクラスタリングしてωグループ（ωは１より大きい整数）とすること、および、学習パターンを分けたωグループの中心を代表パターンとして、クラスタリングによりグループ内部で重複を行って、最良のグループ分けを得ることの各ステップよりなる。



1

【特許請求の範囲】

【請求項１】　１つの文字の他数個の学習パターンをそれぞれ不均等分割すること、
当該文字の学習パターンの特徴を抽出すること、
各学習パターンの特徴から一定数の特徴を選出して、グループを構成すること、
選出した特徴に基づき、学習パターンをクラスタリング法でωグループに分けるが、このωが１より大きい整数であること、
および、
学習パターンをクラスタリングしたωグループの中心を代表パターンとして、クラスタリング法を利用してグループ内に投入して重複させ、最良のグループ分けを実現することという上記のステップを具備する文字認識におけるグループデータベース確立の方法。

【請求項２】　上記クラスタリング法が、ランダムクラスタリング、ＩＳＯＤＡＴＡクラスタリング、ｋ－平均クラスタリングを含み、順序だてて実行されるものである請求項１記載のグループデータベース確立の方法。

【請求項３】　上記ωが、ω≧１５である請求項１記載のグループデータベース確立の方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は文字認識におけるグループデータベース確立の方法に関し、特に文字認識において文字の代表パターンを利用してシステムとしてグループデータベースを確立する方法であって、パターンが多すぎること及び変異が大きすぎることに起因するシステムのグループデータベースの拡張（重複量の増大）を回避すると同時に、システムのデータベースの保存スペースを削減する方法に係る。

【０００２】

【従来の技術】従来において、文字認識の技術分野では、グループデータベース確立の技術について考慮を要する２点、つまりパターン重複量とグループ投入精度とがあった。一般に使用されてきた重複技術では、グループを先ずC_1，C_2，…，C_mと仮設する。文字構成要素（訳注：字素？　次元？）X_iにつき、$1 \leqq i \leqq N$，X_iとC_kとの距離を$d(X_i, C_k) = \min d(X_i, C_k)$としていた。もし、

【０００３】

【数１】

$$X_i \in C_k$$

【０００４】ならば$C_k \cup \{X_i\}$が重複を示す。

【０００５】

【発明が解決しようとする問題点】しかしながら、従来においては、重複量とグループ投入精度とは両立できないものであった。もしグループ投入精度を向上させたいなら重複量を増やさなければならないので、グループ内

2

部に保存できるコードの保存スペースに影響をおよぼすとともに、マッチングに要する時間を増大させるものとなっていた。従って、このような欠点を克服することが、文字認識の技術分野における重要課題となっていた。

【０００６】

【問題点を解決するための手段】上記した公知の問題を解決するために、本発明では、認識プロセスにおいて多数個の候補グループを利用することによって、従来、代表パターンが完全には字形の変異を吸収できなくなり、かえって１グループだけを探し出すために誤識別というエラーを引き起こしていたことを回避するとともに、累積により精度が増大するという現象を利用して、多グループを採用することによりグループ投入精度を向上させる。

【０００７】このような発明の観点から、文字認識におけるグループデータベース確立の方法を提供するもので、下記ステップを具備する。すなわち、１つの文字について多数個の学習パターンをそれぞれ不均等分割すること、学習パターンの特徴を抽出すること、学習パターンの各特徴の中から一定数の特徴を選出して、グループを確立すること、選出された特徴に基づいて、学習パターンをクラスタリングしてωグループ（ωは１より大きい整数）とすること、および学習パターンを分けたωグループの中心を代表パターンとして、クラスタリングによりグループ内部に投入して重複を行って、最良のグループ分けを得ること、という各ステップである。

【０００８】

【実施例】以下、図１のフローチャートを参照しながら、本発明に係るグループデータベース確立の方法を説明する。最初に、ステップＳＴ１において、文字の特徴値を抽出する前に、局部および内部字形のデータを獲得するために、先ず文字画像を分割する。文字の変形、例えば文字の高さや幅の不一致、字形や字画の不均等な分布、ならびに特殊な字画の長さの差異により、もし印刷文字方式の均等分割を行うならば、かならず分割が不正確となって、抽出した特徴値の変異が非常に大きいものとなり、誤識別を発生させる。そこで、本発明は、不均等分割方式を採用する。例をあげるなら、図４の『示』という文字を分割するために、３種類の不均等分割方式、つまり字画密度関数（Ｓｔｒｏｋｅ　Ｄｅｎｓｉｔｙ　Ｆｕｎｃｔｉｏｎ；ＳＤＦ）、周辺面積（Ｐｅｒｉｐｈｅｒａｌ　Ｂａｃｋｇｒｏｕｎｄ　Ａｒｅａ；ＰＢＡ）、周辺輪郭線長さ（Ｃｏｎｔｏｕｒ　Ｌｉｎｅ　Ｌｅｎｇｔｈ；ＣＬＬ）を採用し、それぞれ図５の（Ａ），図５の（Ｂ），図５の（Ｃ）に示す。これら３種類の不均等分割方式は、いずれも公知技術なので、ここでは改めて説明しない。

【０００９】そして、ステップＳＴ２において、本発明に係る実施例は、ＳＤＦ，ＰＢＡ，ＣＬＬという３種類

の不均等分割で６４個の特徴を取り出す。続いて、ステップＳＴ３において、

＊【００１０】

＊【数２】

$$CF_j = \sum \sum \frac{M_{jl} - m_{jk}}{\sigma_{jk}}$$

式中、ｊは特徴の第ｊ次元、ｉ，ｋは第ｉ，ｋの文字、ｍは中心値、

σ_{jk}は文字の第ｊ次元の特徴標準偏差ＳＴＤを表す、

【００１１】を利用して、CF_j 値（ｊ＝１～６４）のなかで最大となる４０個を探して、グループ化された４０次元の特徴とするとともに、残りの２４個を棄却して使用しない。次に、ステップＳＴ４において、文字構成要素のパターンクラスタリングを行う。図２を見ると分かりやすいように、本発明の好適な実施例に係るクラスタリング法（Ｈ－ｃｌａｓｓ）の流れを詳細に示している。本発明の好適な実施例において、２５６８個の文字構成要素を採用している。各文字構成要素に対して、多数個の（本実施例では１００個の）学習パターンをクラスタリング法で先にωグループに分ける。ωは、本実施例では、１５，１７，２０に相当する。クラスタリング法は、まずランダム・クラスタリングを行って、次にイソデータＩＳＯＤＡＴＡクラスタリングを１回、さらにＫ－ｍｅａｎ（Ｋ平均）クラスタリングを行い、後者によって得られた中心値を文字構成要素を代表する代表パターンとする。そして、次の文字構成要素について１０※０個の学習パターンを分解していって、２５６８個の文字構成要素の全部を処理するまで続ける。

【００１２】上記イソデータクラスタリングは、X_1，X_2，…，X_m をC_1，C_2，…，C_m 中に投入して、以下の４ステップを行う必要がある。

（１）X_i を任意にC_k グループ中へ分散して、ｃｏｕｎｔ＝０とする、（２）各X_i につき、$1 \leqq i \leqq N$，$X_i \in C_s$ とし、$d(X_i, C_t) = \min d(X_i, C_k)$，$1 \leqq k \leqq m$とする、（３）もしｔ≠Ｓなら、すなわち$C_t + \{X_i\}$，$C_s = C_s - \{X_i\}$，ｃｏｕｎｔ＝ｃｏｕｎｔ＋１とする、（４）ｉ＝ｉ＋１とし、もしｉ＝ＮなＣｋ 中の成分について改めて中心値を計算し、違うなら（２）に戻る。

【００１３】ここで、

【００１４】

【数３】

$$d(X_j, C_k) \sum \frac{(X_{ij} - C_{kj})^2}{\sigma_{kj} \cdot \sigma_{sj}}$$

特徴値$X_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iF} \end{bmatrix}$、中心値$C_k = \begin{bmatrix} C_{k1} \\ \vdots \\ C_{kF} \end{bmatrix}$、Ｆは特徴の次元数、

$\sigma_k = \begin{bmatrix} \sigma_{k1} \\ \vdots \\ \sigma_{kF} \end{bmatrix}$を第ｋグループの標準偏差ＳＴＤとし、

$\sigma_{sj} = \frac{1}{2568} \sum_{n=1}^{2568} \Delta_{nj}$とするが、

【００１５】Δ_{nj}は、第ｎ字の第ｊ特徴の標準偏差ＳＴＤ（注意；本実施例では、２５６８個の字を使用）とする。また、ｋ－平均クラスタリングは、以下の５ステップを備えている。

（１）X_i を任意にC_k グループ中に分散して入れ、ｃｏｕｎｔ＝０とする、（２）各X_i ごとに、$1 \leqq i \leqq N$，$X_i \in C_s$ とする、（３）もしｔ≠ｓなら$C_t = C_t + \{X_i\}$，$C_s = C_s - \{X_i\}$，ｃｏｕｎｔ＝ｃ★ｏｕｎｔ＋１とする、（４）ｉ＝ｉ＋１とし、C_k 中の成分を利用して改めて中心値を計算する、（５）もしｃｏｕｎｔ＜臨界値Ｔなら停止し、そうでなければｃｏｕｎｔ＝０として、（２）に戻る。

【００１６】２５６８個の文字構成要素の各学習パターンをすべてクラスタリングした後、ステップＳＴ５に進み、システムの代表パターンのクラスタリングを行う。ステップＳＴ５の詳細な流れは図３を参照されたい。こ

5

の時、各パターン毎にｋ－平均クラスタリングでクラス分けしたωグループの中心値を代表パターンとしてグループ内に投入して、もう一度、上記のクラスタリング法を応用して重複させる。そして、パターンを測定して照合するとともに、数個の候補グループを選出する（本実施例では5個としている）。その後、各パターンごとにグループ投入精度が最大となる時のωを取る。そして、もう一度、ステップＳＴ５を繰り返す。本実施例によれ＊

6

＊ば、２５６８字が２００個のパターンを有するシステムにおいて、１００個のパターンを学習パターンとし、他の１００個は測定を行う。本発明により得られる識別効果は極めて顕著なものがあるので、公知技術と比較した表１を参照されたい。

【００１７】

【表１】

	公　知	本　発　明
字数	２５６８	２５６８
クラスタリング法	ｋ－平均	ｋ－平均
学習パターン数	１００パターン／字	１００パターン／字
重複方法	パターンをグループに投入	パターンクラスタリンググループ中心をグループに投入
重複量	１０	２．８３
候補グループ数	１	５
測定パターン数	１００	１００
各グループの字数平均	７００	７６０
グループ投入精度	９７．４６％	９９．５８％

【００１８】上の表から分かるように、本発明はグループ中心をグループに投入することによって、重複量を削減させるばかりではなく、グループ投入精度を向上させる。以上に記載した比較的好ましい実施例は、本発明を説明するためのもので、本発明を制限するものではない。この技術分野に詳しい当業者なら多くの修正とバリエーションとが可能であるが、それでもなお、以下の特許請求の範囲で限定した発明のカテゴリーと精神とを回避できるものではない。

【図面の簡単な説明】

【図１】本発明に係るグループデータベース確立の方法を示すフローチャートである。

【図２】図１のステップＳＴ４の詳細な流れを示すフロ※ーチャートである。

※

【図３】図１のステップＳＴ５の詳細な流れを示すフローチャートである。

【図４】１つの中国文字『示』の不均等分割を示す図である。

【図５】図４の『示』という文字を３種類の不均等分割方式で分割して、特徴抽出した図である。

【符号の説明】

ＳＴ１　不均等分割

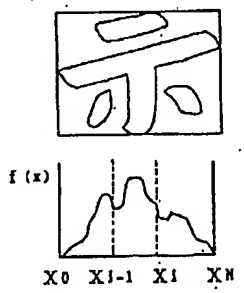
ＳＴ２　特徴抽出

ＳＴ３　特徴選出

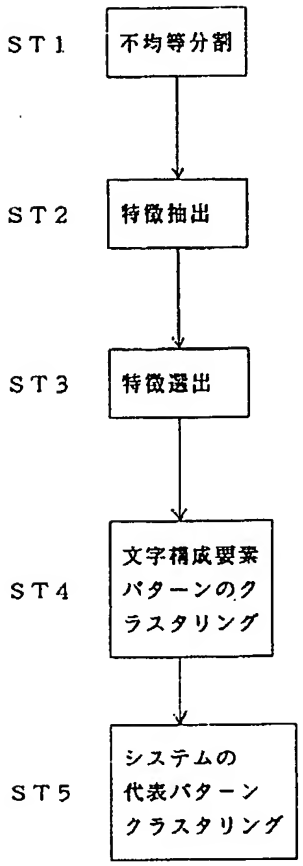
ＳＴ４　文字構成要素のクラスタリング

ＳＴ５　システムの代表パターンクラスタリング

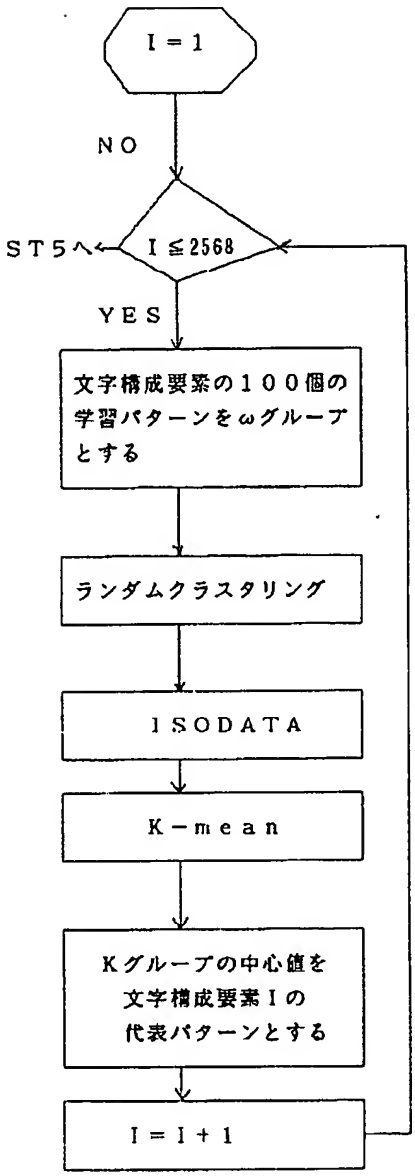
【図４】



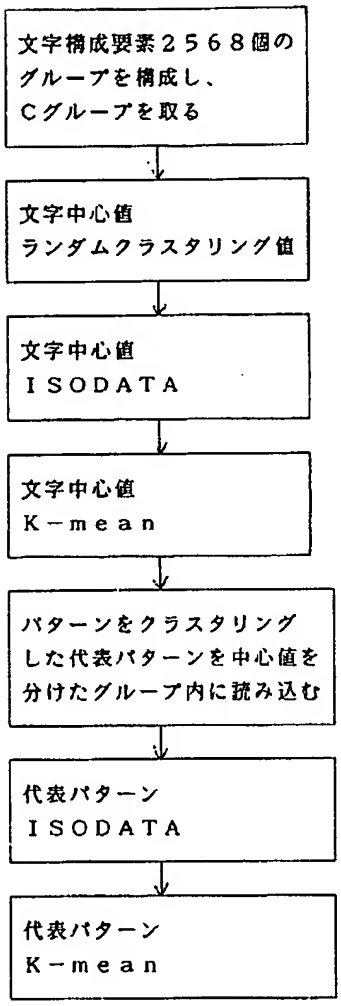
【図1】


ST1
不均等分割
ST2
特徴抽出
ST3
特徴選出
ST4
文字構成要素
パターンのク
ラスタリング
ST5
システムの
代表パターン
クラスタリング

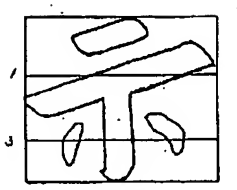
【図2】


I＝1
NO
ST5へ
I≦2568
YES
文字構成要素の１００個の
学習パターンをωグループ
とする
ランダムクラスタリング
ISODATA
K－mean
Kグループの中心値を
文字構成要素Iの
代表パターンとする
I＝I＋1

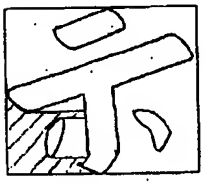
【図3】


文字構成要素２５６８個の
グループを構成し、
Cグループを取る
文字中心値
ランダムクラスタリング値
文字中心値
ＩＳＯＤＡＴＡ
文字中心値
Ｋ－ｍｅａｎ
パターンをクラスタリング
した代表パターンを中心値を
分けたグループ内に読み込む
代表パターン
ＩＳＯＤＡＴＡ
代表パターン
Ｋ－ｍｅａｎ

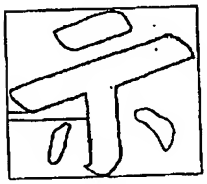
【図5】



(A)



(B)



(C)

フロントページの続き

(72)発明者　林　文　▲ぶん▼
台湾台北市北投區温泉里10鄰温泉路58巻9
－1號